# Exemple PLS avec SAS

This example, from Umetrics (1995), demonstrates different ways to examine a PLS model. The data come from the field of drug discovery. New drugs are developed from chemicals that are biologically active. Testing a compound for biological activity is an expensive procedure, so it is useful to be able to predict biological activity from cheaper chemical measurements. In fact, computational chemistry makes it possible to calculate certain chemical measurements without even making the compound. These measurements include size, lipophilicity, and polarity at various sites on the molecule. The following statements create a data set named *pentaTrain*, which contains these data.

You would like to study the relationship between these measurements and the activity of the compound, represented by the logarithm of the relative Bradykinin activating activity (*log_RAI*). Notice that these data consist of many predictors relative to the number of observations. Partial least squares is especially appropriate in this situation as a useful tool for finding a few underlying predictive factors that account for most of the variation in the response. Typically, the model is fit for part of the data (the "training" or "work" set), and the quality of the fit is judged by how well it predicts the other part of the data (the "test" or "prediction" set). For this example, the first 15 observations serve as the training set and the rest constitute the test set (refer to Ufkes et al. 1978; Ufkes et al. 1982).

When you fit a PLS model, you hope to find a few PLS factors that explain most of the variation in both predictors and responses. Factors that explain response variation provide good predictive models for new responses, and factors that explain predictor variation are well represented by the observed values of the predictors. The following statements fit a PLS model with two factors and save predicted values, residuals, and other information for each data point in a data set named *outpls*.

**PROGRAMME SAS**

```
options ls=64 ps=80 nodate nonumber;
data pentaTrain;
      input obsnam $ S1 L1 P1 S2 L2 P2
                     S3 L3 P3 S4 L4 P4
                     S5 L5 P5  log_RAI @@;
      n = _n_;
      datalines;
   VESSK    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
             1.9607 -1.6324  0.5746  1.9607 -1.6324  0.5746
             2.8369  1.4092 -3.1398                   0.00
   VESAK    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
             1.9607 -1.6324  0.5746  0.0744 -1.7333  0.0902
             2.8369  1.4092 -3.1398                   0.28
   VEASK    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
             0.0744 -1.7333  0.0902  1.9607 -1.6324  0.5746
             2.8369  1.4092 -3.1398                   0.20
   VEAAK    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
             0.0744 -1.7333  0.0902  0.0744 -1.7333  0.0902
             2.8369  1.4092 -3.1398                   0.51
   VKAAK    -2.6931 -2.5271 -1.2871  2.8369  1.4092 -3.1398
             0.0744 -1.7333  0.0902  0.0744 -1.7333  0.0902
             2.8369  1.4092 -3.1398                   0.11
   VEWAK    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
            -4.7548  3.6521  0.8524  0.0744 -1.7333  0.0902
             2.8369  1.4092 -3.1398                   2.73
   VEAAP    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
             0.0744 -1.7333  0.0902  0.0744 -1.7333  0.0902
            -1.2201  0.8829  2.2253                   0.18
   VEHAK    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
             2.4064  1.7438  1.1057  0.0744 -1.7333  0.0902
             2.8369  1.4092 -3.1398                   1.53
   VAAAK    -2.6931 -2.5271 -1.2871  0.0744 -1.7333  0.0902
             0.0744 -1.7333  0.0902  0.0744 -1.7333  0.0902
             2.8369  1.4092 -3.1398                  -0.10
   GEAAK     2.2261 -5.3648  0.3049  3.0777  0.3891 -0.0701
             0.0744 -1.7333  0.0902  0.0744 -1.7333  0.0902
             2.8369  1.4092 -3.1398                  -0.52
   LEAAK    -4.1921 -1.0285 -0.9801  3.0777  0.3891 -0.0701
             0.0744 -1.7333  0.0902  0.0744 -1.7333  0.0902
             2.8369  1.4092 -3.1398                   0.40
   FEAAK    -4.9217  1.2977  0.4473  3.0777  0.3891 -0.0701
             0.0744 -1.7333  0.0902  0.0744 -1.7333  0.0902
             2.8369  1.4092 -3.1398                   0.30
   VEGGK    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
             2.2261 -5.3648  0.3049  2.2261 -5.3648  0.3049
             2.8369  1.4092 -3.1398                  -1.00
   VEFAK    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
            -4.9217  1.2977  0.4473  0.0744 -1.7333  0.0902
             2.8369  1.4092 -3.1398                   1.57
   VELAK    -2.6931 -2.5271 -1.2871  3.0777  0.3891 -0.0701
            -4.1921 -1.0285 -0.9801  0.0744 -1.7333  0.0902
             2.8369  1.4092 -3.1398                   0.59
   ;
   proc reg data=pentaTrain;
      model log_RAI = S1-S5 L1-L5 P1-P5;
   run;
ods graphics on;

   proc pls data=pentaTrain;
      model log_RAI = S1-S5 L1-L5 P1-P5;
   run;
    proc pls data=pentaTrain nfac=2 plot=(ParmProfiles VIP);
      model log_RAI = S1-S5 L1-L5 P1-P5;
   run;
 ods graphics off;
```

Dans cette première partie la procédure REG est utilisée. Nous constatons un certain nombre de problèmes dus au fait de disposer de seulement de 15 observations pour un modèle construit à partir de 15 variables explicatives :

**TABLEAU 1 : Procédure REG**

Modèle : MODEL1
Variable dépendante : log_RAI

| | |
|---|---|
| Nombre d'observations lues | 15 |
| Nombre d'observations utilisées | 15 |

Analyse de variance

| Source | DDL | Somme des carrés | Moyenne quadratique |
|---|---|---|---|
| Modèle | 11 | 11.39423 | 1.03584 |
| Erreur | 3 | 0.10141 | 0.03380 |
| Total sommes corrigées | 14 | 11.49564 | |

Analyse de variance

| Source | Valeur F | Pr > F |
|---|---|---|
| Modèle | 30.64 | 0.0084 |
| Erreur | | |
| Total sommes corrigées | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.18385 | R carré | 0.9912 |
| Moyenne dépendante | 0.45200 | R car. ajust. | 0.9588 |
| Coeff Var | 40.67543 | | |

NOTE: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

NOTE: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

```
L5 =  1.04118 * Intercept + 0.12973 * S5
P2 =  -4.53088 * Intercept + 1.7768 * S2 - 2.58979 * L2
P4 =  0.22776 * Intercept + 0.25198 * S4 + 0.09018 * L4
P5 =  0.6118 * Intercept - 1.32243 * S5
```

Valeurs estimées des paramètres

| Variable | DDL | Valeur estimée des paramètres | Erreur type | Valeur du test t | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | B | 0.44572 | 4.15207 | 0.11 | 0.9213 |
| S1 | 1 | 0.12634 | 1.05805 | 0.12 | 0.9125 |
| S2 | B | 0.24757 | 0.15668 | 1.58 | 0.2122 |
| S3 | 1 | -0.04959 | 0.04761 | -1.04 | 0.3741 |
| S4 | B | -0.11379 | 0.09455 | -1.20 | 0.3151 |
| S5 | B | 0.03536 | 0.05320 | 0.66 | 0.5538 |
| L1 | 1 | 0.26808 | 1.17196 | 0.23 | 0.8338 |

Procédure REG
Modèle : MODEL1
Variable dépendante : log_RAI

Valeurs estimées des paramètres

| Variable | DDL | Valeur estimée des paramètres | Erreur type | Valeur du test t | Pr > \|t\| |
|---|---|---|---|---|---|
| L2 | B | -0.15081 | 0.17790 | -0.85 | 0.4589 |
| L3 | 1 | 0.36201 | 0.08500 | 4.26 | 0.0237 |
| L4 | B | -0.08741 | 0.09185 | -0.95 | 0.4114 |
| L5 | 0 | 0 | . | . | . |
| P1 | 1 | -0.44236 | 1.22959 | -0.36 | 0.7429 |
| P2 | 0 | 0 | . | . | . |
| P3 | 1 | 0.11778 | 0.24916 | 0.47 | 0.6687 |
| P4 | 0 | 0 | . | . | . |
| P5 | 0 | 0 | . | . | . |

**TABLEAU 2: The PLS Procedure**

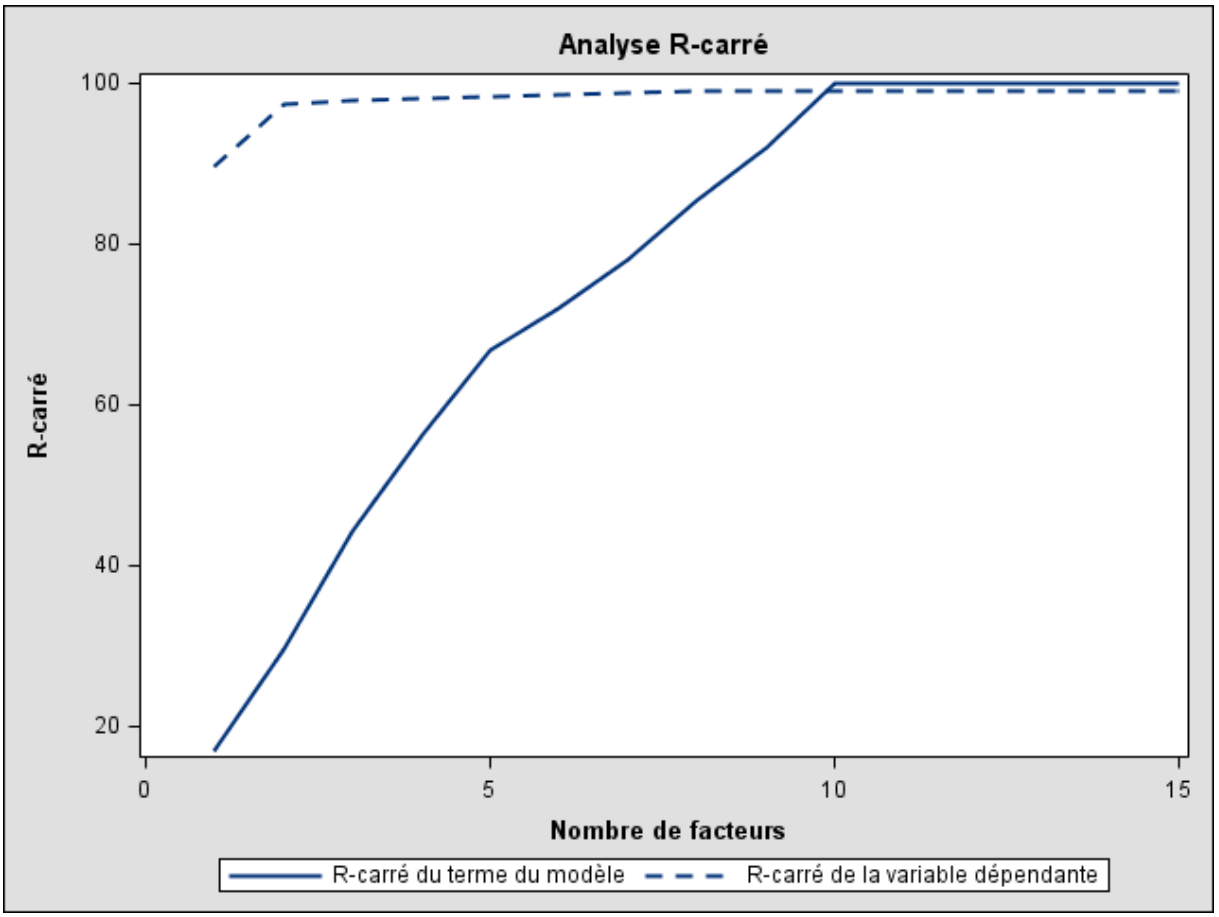
| | |
|---|---|
| Data Set | WORK.PENTATRAIN |
| Factor Extraction Method | Partial Least Squares |
| PLS Algorithm | NIPALS |
| Number of Response Variables | 1 |
| Number of Predictor Parameters | 15 |
| Missing Value Handling | Exclude |
| Number of Factors | 15 |

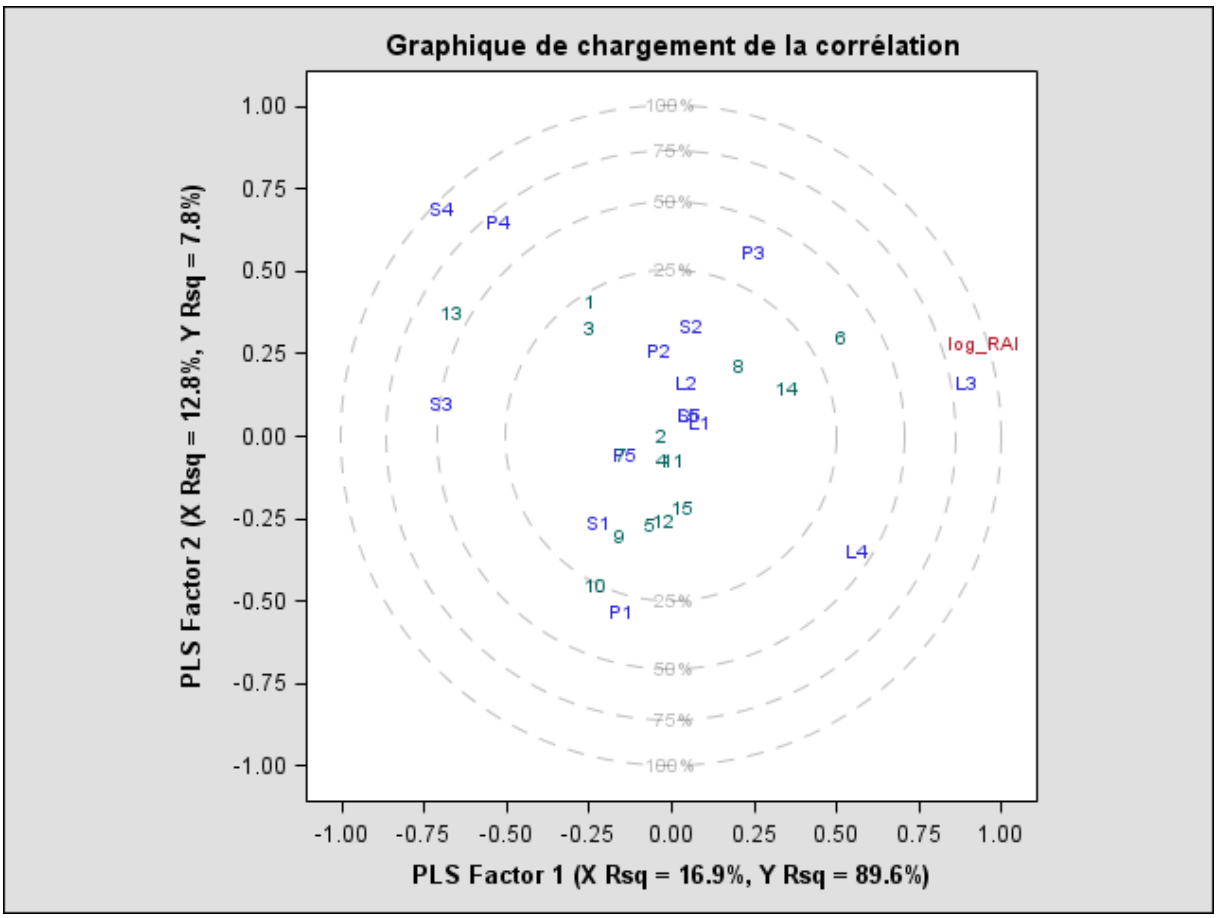
| | |
|---|---|
| Number of Observations Read | 15 |
| Number of Observations Used | 15 |

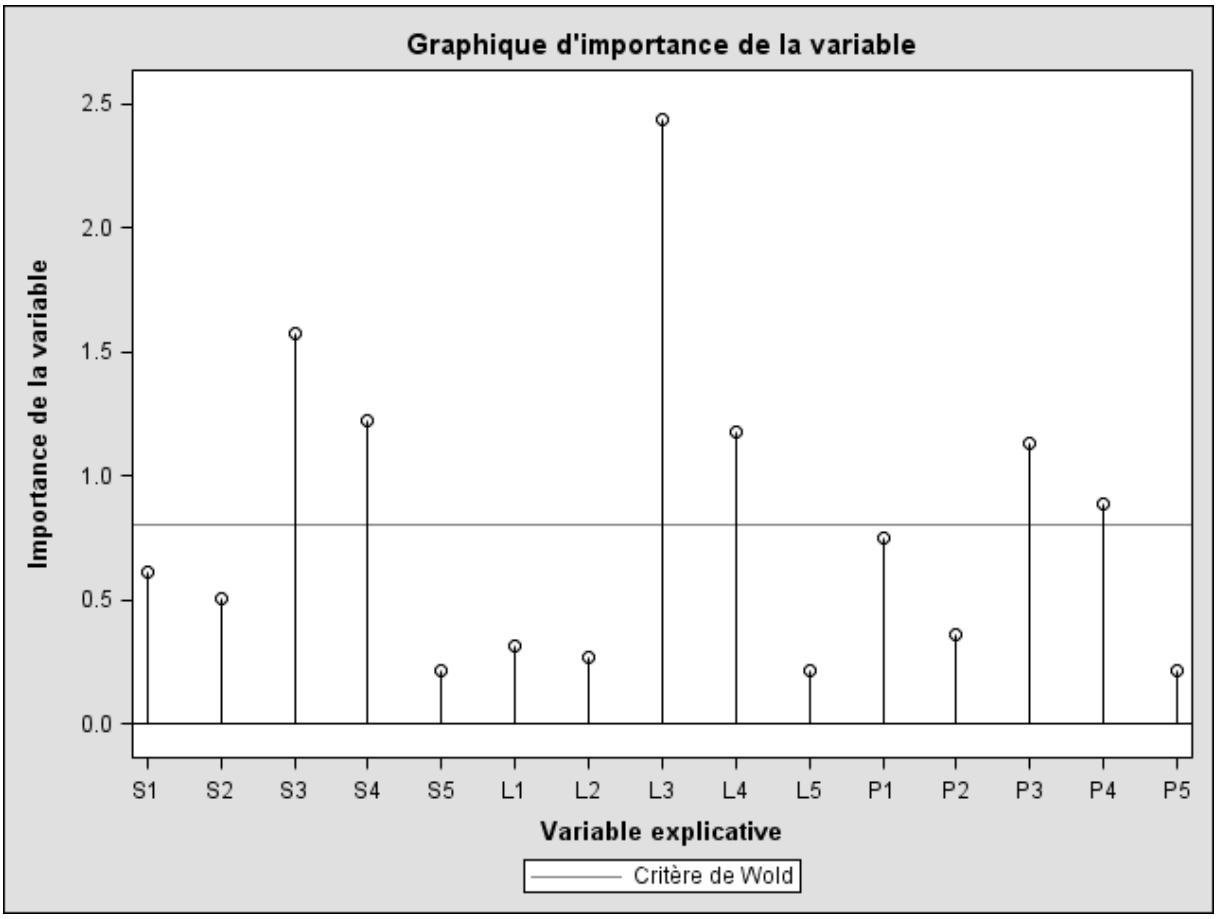
On constate que la première composante PLS explique 16,9% de la variation des 15 prédicteurs mais dispose d'un R2 de 89,63% avec la variable à expliquer.
Les deux premières composantes PLS expliquent 29,67% de la variation des 15 prédicteurs, pour un R2 avec la variable à expliquer de 97,47%. Les valeurs affichées dans le tableau ci-dessous sont illustrées par le graphique qui suit.

**Variation en pourcentage expliquée par Partial Least Squares Factors**

| Nombre de facteurs extraits | Effets du modèle | | Variables dépendantes | |
|---|---|---|---|---|
| | Actuel | Total | Actuel | Total |
| 1 | 16.9014 | 16.9014 | 89.6399 | 89.6399 |
| 2 | 12.7721 | 29.6735 | 7.8368 | 97.4767 |
| 3 | 14.6554 | 44.3289 | 0.4636 | 97.9403 |
| 4 | 11.8421 | 56.1710 | 0.2485 | 98.1889 |
| 5 | 10.5894 | 66.7605 | 0.1494 | 98.3383 |
| 6 | 5.1876 | 71.9481 | 0.2617 | 98.6001 |
| 7 | 6.1873 | 78.1354 | 0.2428 | 98.8428 |
| 8 | 7.2252 | 85.3606 | 0.1926 | 99.0354 |
| 9 | 6.7285 | 92.0891 | 0.0725 | 99.1080 |
| 10 | 7.9076 | 99.9967 | 0.0000 | 99.1080 |
| 11 | 0.0033 | 100.0000 | 0.0099 | 99.1179 |
| 12 | 0.0000 | 100.0000 | 0.0000 | 99.1179 |
| 13 | 0.0000 | 100.0000 | 0.0000 | 99.1179 |
| 14 | 0.0000 | 100.0000 | 0.0000 | 99.1179 |
| 15 | 0.0000 | 100.0000 | 0.0000 | 99.1179 |


Analyse R-carré
R-carré
100
80
60
40
20
0
5
10
15
Nombre de facteurs
R-carré du terme du modèle
R-carré de la variable dépendante


Graphique de chargement de la corrélation
PLS Factor 2 (X Rsq = 12.8%, Y Rsq = 7.8%)
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
100%
75%
50%
25%
S4
P4
P3
13
1
3
S2
P2
6
log_RAI
8
L2
14
L3
S3
L1
2
P5
11
S1
9
12
15
L4
10
P1
-1.00
-0.75
-0.50
-0.25
0.00
0.25
0.50
0.75
1.00
PLS Factor 1 (X Rsq = 16.9%, Y Rsq = 89.6%)


Graphique d'importance de la variable
Importance de la variable
2.5
2.0
1.5
1.0
0.5
0.0
S1
S2
S3
S4
S5
L1
L2
L3
L4
L5
P1
P2
P3
P4
P5
Variable explicative
Critère de Wold